IDIAP RESEARCH REPORT

INFORMATION THEORETIC ANALYSIS OF PRODUCTION-PERCEPTION EFFICIENCY: CASE STUDY OF SPEECH PATHOLOGY

Afsaneh Asaei Milos Cernak Hervé Bourlard

Idiap-RR-30-2016

DECEMBER 2016

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

Information Transmission Analysis of Production-Perception Efficiency: Case Study of Speech Pathology

Afsaneh Asaei, *Senior Member, IEEE,* Milos Cernak, *Member, IEEE,*
Hervé Bourlard, *Fellow, IEEE,*

Abstract

Phonological classes define articulatory-free and articulatory-bound phone attributes. Deep neural network is used to estimate the probability of phonological classes from the speech signal. In theory, a unique combination of phone attributes form a phoneme identity. Probabilistic inference of phonological classes thus enables estimation of their compositional phoneme probabilities. A novel information theoretic framework is devised to quantify the information conveyed by each phone attribute, and assess the speech production quality for perception of phonemes. As a use case, we hypothesize that disruption in speech production leads to information loss in phone attributes, and thus confusion in phoneme identification. We quantify the amount of information loss due to dysarthric articulation available in the TORGO database. A novel information measure is formulated to evaluate the deviation from an ideal phone attribute production leading us to distinguish healthy production from pathological speech.

Index Terms

Information transmission, Speech production, Speech perception, Motor speech disorders

I. Introduction

Invariant speech representation is fundamental for speech modeling and classification. In this context, phonetic and phonological representations are widely regarded as robust representations invariant to

Authors are with Idiap Research Institute, Centre du Parc, Rue Marconi 19, 1920 Martigny, Switzerland. Hervé Bourlard is also affiliated with École Polytechnique Fédérale de Lausanne, Switzerland. Emails: afsaneh.asaei@idiap.ch, milos.cernak@idiap.ch, herve.bourlard@idiap.ch.

speaker and acoustic conditions. These representations are also supported by psycho- and neuro- linguistic studies of speech production and perception. The present paper proposes an information theoretic analysis of phonetic and phonological representations. We are interested is assessment of speech production quality and perception. A schematic functional view of production-perception processes is illustrated in Fig. 1.

Speech production is one of the most complex motor coordination processes of human brain. It involves a networked system of brain areas that each contribute in unique ways [1]. A simplified psycholinguistic model of speech production [2], [3] typically consists of linguistic, motor planning and motor programming/execution stages. The linguistic stage is characterized by phonological encoding, namely the preparation of an abstract speech code. Speech code is an invariant speech representation that lies in the intersection of the cognitive and motor control processes.

Speech code is greatly debated in motor control, psycholinguistics, neuropsychology and speech neuroscience. Recent findings suggest that speech code includes articulatory gestures [4]–[7], and auditory and somatosensory targets [8]. Speech code can be defined at phonetic or phonological levels. In the present study, we assume that the invariant speech code is defined by composition of phonological classes. The phonological classes refer to articulatory-free and articulatory-bound phone attributes, and they are correlated with the auditory and acoustic events [9]. Exploiting phone attributes facilitates development of our theoretical framework for analysis of speech production and perception. This framework can be applied for alternative representations.

Speech perception refers to the mapping from sound to the internal linguistic representation. Earlier studies are conducted in the context of syllable recognition and investigate its relation to the mechanism of auditory processing. Pioneering work of Harvey Fletcher demonstrated that human recognition acts on the principle of processing parallels of independent streams enabling partial recognition and merging of the independent evidences for speech recognition [10], [11]. Although Fletcher established his work for processing of disjoint frequency ranges (auditory events), the notion of independent processing influenced later development of speech perception theories regardless of auditory processing [12].

An important perspective to speech perception relies on inverse production processing or phonological decoding. The decoding process is quite complex and a complete explanation of how humans recognize syllables and phonemes remains elusive [12]. In this context, the motor theory is probably one of the oldest that has been re-investigated and revised extensively [12]–[14]. According to the motor theory of speech perception, the objects of speech perception are articulatory rather than acoustic or auditory events [12], [15]. Although this theory has been partially controversial, several experimental evidence support the idea that perception operates on the principle of detecting the underlying structures or articulatory gestures [12],

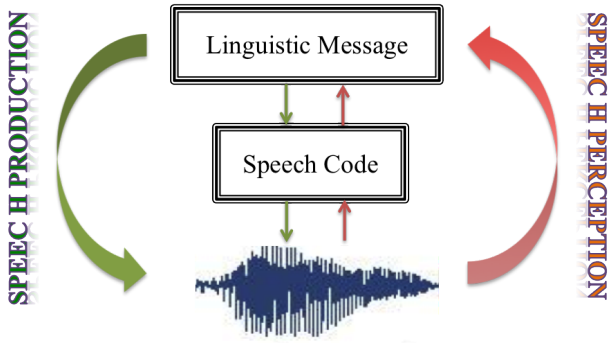

Fig. 1. A schematic functional view of speech production and perception.

[14], [16]. The vocal tract actions (e.g., the closing and opening of the lips during the production of /pa/) structure the acoustic signal. As noted in [14], "speakers produce phonetic frames as individual or as coupled gestures of the vocal tract. The gestures cause information in acoustic speech signals for the segmental structure of utterances, and that experienced listeners are sensitive to that information".

The psycholinguistic theories assert that a unique binary mapping exists between phonemes and phonological classes, and speech can be seen as the molecules of alphabetic atoms [14], [17]. However, accessing the compositional atoms from the speech signal is an open problem. In practice, speech manifests itself in continuous forms that may be attributed to multiple classes. A great challenge in this context is pertained to speech coarticulation and supra-segmental variations [16].

The present study builds on the success of deep neural network (DNN) in estimation of class-conditional posterior probabilities. We apply DNNs for probabilistic characterization of the phonological classes [18], [19]. We advance the linguistic binary association of the phoneme and phonological classes by considering the dynamic probabilistic associations adapting to the production condition. We define phonological compositions as the set of phonological classes forming the phoneme identities.

We consider the linguistic message being present in form of phoneme transcription. The production machinery is then regarded as a channel that transmit the phoneme information to phonological classes or phone attributes. Accordingly, the phoneme perception operates on the principle of phonological class inference and composition for phoneme identification. DNN estimates the phonological class probabilities from the speech signal. In an ideal speech production condition, high probabilities are estimated whereas the disruption in production results in small probabilities. We propose an information theoretic approach to quantify the information content of phone attributes.

As a case study, we exploit the proposed method in the context of production assessment in speech

pathology. This enables us to contrast control/healthy and pathological speech to reveal the degree of information loss apparent at the level of individual phone attributes. Considering phonemes as composition of phone attributes, the most informative attributes for phoneme identification are determined. Moreover, the phonemes mainly affected by production impairment are identified and their information loss is quantified. We measure the degree of impairment or deviation from an ideal production that enables us to distinguish healthy speech from impaired production.

The rest of the paper is organized as follows. The framework for estimation of phonological class probabilities is outlined in Section II. We explain the information theoretic method for assessment of speech production in terms of phone attribute information in Section III. The measures of information loss are formulated in Section IV. The numerical results are evaluated in Section V, and finally the concluding remarks are drawn in Section VI.

II. Phonological Posteriors

We use DNN for estimation of class-conditional posterior probabilities [20]. In this framework, K independent DNNs take as input acoustic features derived from short frames of speech signal, and estimate the class-conditional posterior probability of K phonological classes given the input acoustic features. The DNN output probabilities are briefly dubbed *phonological posteriors*. Each component of the phonological posterior represents the probability of a phone attribute in the speech signal. These attributes describe speech segments phonemes using binary labels; for example, phonological classes of [consonantal], [anterior], [voice] and [nasal] compose phoneme /M/ [20].

Linguistics define two traditional speech structures: (i) cognitive structures represented by (discrete) canonical representation, and (ii) surface structures exhibited by (continuous) observed representation patterns. The phonological posteriors are associated with the surface structures. The phonological posteriors yield a parametric speech representation, and the trajectories of the articulatory-bound phonological posteriors correspond to the distal representation of the gestures in the gestural model of speech production (and perception). Hence, we hypothesize that they represent the probabilistic relation of the canonical phonetic and phonological classes to a distal representation of the (co-articulated) speech code.

The present study exploits phonological posteriors as essential representations to quantify the information content of produced phone attributes and the information loss due to impaired speech production. To that end, we use information theory for transmission analysis of the production channel as explained in the following Section III.

III. Information Transmission Analysis

In this section, we formulate an information theoretic analysis of speech production and perception. The proposed approach builds on the seminal work of Miller and Nicely on analysis of perceptual confusion and information loss in noisy communication systems [21]. The original theory of information transmission analysis (ITA) is developed for quantification of information conveyed by binary phone attributes, such as voicing, place and manner of articulation [21].

In practice, however, co-articulation and supra-segmental variations such as stress affect the binary association between phoneme and phonological classes [22], [23]. Therefore, the present paper adopts the probabilistic estimation of phone attributes for ITA. DNN provides the phonological posteriors that quantifies the precision of phonetic attributes detected from the speech signal.

A. Production of Phone Attributes

The following production scenario is considered. A phonetic transcription is provided, which is encoded through the speech production process in terms of phone attributes as depicted in Fig. 2. A listener (judging the speech production quality) may detect/infer phone attributes towards recognition of the speech signal.

Fig. 2. Phone attribute encoding: Speech production channel transmits the *source* phoneme information through production of the *target* phone attributes.

It may be noted that the source information can be presented in a larger granularity such as syllables or words, and the target of speech production can be considered different than phone attributes such as neuromuscular commands. The scenario hypothesized here (Fig. 2) facilitates derivation of our analysis. Nevertheless, the theory and algorithm remain applicable for different granularity of source and target units.

We exploit information theory to quantify the information content of phone attributes to convey phoneme transcription.

B. ITA of Binary Phonetic-Phonological Association

The analysis is based on the mathematical theory developed by Claude Shannon [24] to calculate the information quantity transmitted over a noisy channel. This theory is built on the fundamental measure

of information known as the *Shannon information index* or *entropy*. Shannon proposed this measure to quantify the information content or entropy (uncertainty) in strings of text. The idea was that the more different letters there are, and the more equal their proportional abundances in the string of interest, the more difficult it is to correctly predict which letter will be the next one in the string.

To apply ITA on binary phonetic-phonological association, we define random variables corresponding to phoneme categories and phonological classes.

The random variable denoting phoneme categories is an L dimensional random variable S with categorical distribution $(p_{s_1}, \ldots, p_{s_L})$ where p_{s_l} denotes the probability of phoneme s_l. This random variable corresponds to the source input of the speech production channel (c.f. Fig. 2).

At the output, K phonological classes are the targets constituting the set of $Q = \{q_1, \ldots, q_K\}$ where every phonological class q_k is a discrete random variable taking binary values $\{0, 1\}$, with probability $p(q_k = 1) = p_{q_k}$. The speech production channel is characterized by the joint probabilities $\{p(q_1, S), \ldots, p(q_K, S)\}$.

The goal of applying ITA on binary phonetic-phonological association is to quantify the information content of every individual phone attribute in phoneme transcription. This procedure relies on two quantities as explained below.

1. Source information: The quantity $H(S)$ measures the amount of information made available to the speech production channel by the phonetic transcription S. It is calculated based on the definition of entropy for categorical random variables expressed as

$$\mathcal{H}_{\text{source}} = H(S) = \sum_{l=1}^{L} H(s_l), \quad \text{where} \tag{1}$$

$$H(s_l) = -p_{s_l} \log_2 p_{s_l}. \tag{2}$$

Speech production transmits this information through phone attributes, and accordingly, the perception relies on inference of the compositional phonological classes for phoneme identification (more details in Section IV).

2. Transmitted information: The quantity of information transmitted by the production channel amounts to the mutual information between phonological classes and phonetic transcription.

In former psycholinguistic studies, the phone attributes are defined as binary variables. Hence, the information of an individual phonological class $q_k, \forall q_k \in Q$ is calculated as

$$H(q_k) = -p_{q_k} \log_2 p_{q_k} - (1 - p_{q_k}) \log_2 (1 - p_{q_k}) \tag{3}$$

Algorithm 1 ITA of Phonetic-Phonological Mapping

Input: Table of binary phonetic-phonological association. Phonetic transcription of the data.

Output: Information content of phonetic transcription and phonological classes.

1) Construct matrix $M_{K\times L}$ such that every component M_{kl} is 0 when the phonetic attribute k is missing in phoneme l, and 1, otherwise.

2) Count the number of times each phoneme is present at the phonetic transcription to form vector $N = [n_{s_1} \dots n_{s_l}]^\top$.

3) $p(q_k, s_l)$: Convert the frequency matrix $F = MN$, to joint probability matrix through normalization $P = F/C$.

4) $p(q_k)$: Obtain phonological probabilities via summation of columns of P (marginalization over phonemes).

4) $p(s_l)$: Obtain phoneme probabilities via summation of rows of P (marginalization over phonological classes).

Return $H(S)$ using (1)-(2) and $I(q_k, S)$ using (3)-(5).

Given the phoneme transcription, the information of phonological classes is obtained as

$$\begin{aligned} H(q_k|S) &= -\sum_{l=1}^{L} p(q_k, s_l) \log_2 p(q_k|s_l) \\ &= -\sum_{l=1}^{L} p(q_k, s_l) \log_2 \frac{p(q_k, s_l)}{p(s_l)} \end{aligned} \tag{4}$$

The mutual information quantifies the amount of uncertainty resolved by a phone attribute, thus calculated as

$$\mathcal{I}^k_{\text{transm-binary}} = I(q_k, S) = H(q_k) - H(q_k|S) \tag{5}$$

The quantity $I(q_k, S)$ measures the amount of information made available by the speech production channel to the listener through phonological class q_k.

To implement the binary ITA, the table of phonetic-phonological mapping and the phonetic transcription of the data are required. The probability of every phone attribute being present can be obtained by frequency approach based on counting and relative ratios. The summary of this procedure is outlined in Algorithm 1.

The limitation of the binary association is that it requires detection of the attributes by human subjects, and measurement of the degree an attribute is present is not feasible [21], [25]. In contrast to the binary mapping, in practice a complex function governs the phonetic-phonological association that motivates the use of advanced computational methods for probabilistic characterization. The attributes can be produced with some precision, where high precision leads to higher amount of information content. The low-precision indicates that the attribute may contribute less in resolving the confusion between multiple

phoneme identities. In the next Section III-C, we will see how application of ITA on probabilistic association of phonetic-phonological classes obtained from DNN enables a more practical information transmission analysis.

C. ITA of Probabilistic Phonetic-Phonological Association

The probabilistic association is obtained from DNN phonological posteriors. Application of DNNs enables a computerized method of quantifying the accuracy of phone attribute production, that can be further employed in assessment of speech production quality.

We define z_t as the random variable which can take values of the set of phonological classes $Q = \{q_1, \ldots, q_K\}$; t indexes the time frame. The probabilities of all phonological classes $\{p(z_t = q_1|x_t), \ldots, p(z_t = q_K|x_t)\}$ are estimated by K DNNs [20] each specifically trained to detect one of the classes from the input acoustic speech feature x_t.

The amount of information transmitted by the speech production channel is estimated as the multivariate mutual information [26] $I(S, q_k, z_t)$ between the phonetic transcript, the binary associated phonological class and the probabilistic presence of all phonological classes as follows

$$\begin{aligned}\mathcal{I}^k_{\text{transm-posteriors}} =& I(S, q_k, z_t), \quad \forall k \in \{1, \ldots, K\} \\ =& H(S, q_k, z_t) - H(S, q_k) - H(q_k, z_t) \\ & - H(S, z_t) + H(q_k) + H(S) + H(z_t)\end{aligned} \tag{6}$$

where

$$H(S, q_k, z_t) = -\sum_{l=1}^{L} p(q_k, s_l, z_t) \log_2 p(q_k, s_l, z_t) \tag{6a}$$

$$H(S, q_k) = -\sum_{l=1}^{L} p(q_k, s_l) \log_2 p(q_k, s_l) \tag{6b}$$

$$H(q_k, z_t) = -p(q_k, z_t) \log_2 p(q_k, z_t) \tag{6c}$$

$$H(S, z_t) = -\sum_{l=1}^{L} p(s_l, z_t) \log_2 p(s_l, z_t) \tag{6d}$$

$$H(z_t) = -p(z_t) \log_2 p(z_t) dz \tag{6e}$$

To implement this procedure, the DNN phonological posteriors are used as follows. If the acoustic frame x_t is the result of the production of phone attribute q_k, we assume that $p(x_t|z_t, q_k) = p(x_t|q_k)$; the intuition is that the physical process leading to the production of x_t is guided by q_k (the linguistic code) and the variable z_t is an abstract notion to exploit probabilistic association of the DNN to all

phonological classes. Hence, given the physical state of q_k, the observation x_t is independent of z_t or by Bayes theorem $p(z_t|q_k, x_t) = p(z_t|q_k)$. Similarly, if we consider the production of x_t associated to the phoneme s_l, the DNN output phonological posteriors yields $p(z_t|s_l)$. Thereby, the joint probabilities required to calculate (6) are estimated through conditional probabilities as $p(q_k, z_t) = p(z_t|q_k)p(q_k)$ and $p(s_l, z_t) = p(z_t|s_l)p(s_l)$ where $p(q_k)$ and $p(s_l)$ are known from phonetic transcription, and $p(z_t|q_k)$ and $p(z_t|s_l)$ are directly available from the phonological posteriors.

In general, the multivariate mutual information for three variables can be positive or negative [26]. The positive value indicates a redundancy. In our analysis of the transmitted information, $I(S, q_k, z_t)$ is expected to be positive for all phonological classes. This expectation is due to the redundancy observed at the level of auditory and cortical processes involved in speech production and perception [27], [28]. The redundancy is further analyzed in the following Section III-D.

D. Redundancy in Production of Phonemes

A composition of multiple phonological classes form a phoneme identity. To quantify the amount of redundancy pertained to the phonological compositions, we consider a phoneme s_l composed of K_l phonological classes. The compositional redundancy can then be obtained as the difference between constituting phonological information and the phoneme information expressed as

$$\mathcal{R}^l_{\text{phoneme}} = \sum_{k=1}^{K_l} \mathcal{I}^k_{\text{transm-}} - H(s_l) \tag{7}$$

where $\mathcal{I}^k_{\text{transm-}}$ may be calculated from either binary or probabilistic phonetic-phonological association defined in (5) or (6) respectively; $H(s_l)$ is defined in (2).

The production channel capacity indicates the maximum amount of information that can be transmitted if no error occurs. This ideal situation corresponds to the binary phonetic-phonological association. In this case, $H(q_k|S) = 0$ (4), and the capacity amounts to the overall transmitted information $\mathcal{I}_{\text{transm-binary}} = \sum_{k=1}^{K} \mathcal{I}^k_{\text{transm-binary}}$ where $\mathcal{I}^k_{\text{transm-binary}}$ has the maximum value $H(q_k)$.

Considering the binary association, the theoretical redundancy is obtained, whereas exploiting the probabilistic association yields an actual redundancy present for perception of phonemes as a composition of phone attributes. We evaluate this redundancy in Section V, and study the implications for perceptual loss of phoneme information. In the following Section IV, the information loss objective measures are derived.

IV. Information Loss

The proposed information theoretic analysis of probabilistic phonetic-phonological association enables us to quantify the amount of information conveyed by an individual phone attribute. As a use case, we calculate the information for healthy and impaired speech production, and measure the amount of phonological and phonetic information loss due to production disruption. This idea leads to formulation of a novel compositional information index to assess the production fluency relying on probabilistic estimation of phone attributes.

A. Phonological Information Loss

We compare two information quantities obtained from healthy speech production and impaired production. The difference measures to what extent each of the phone attributes has been disrupted. To state in formally, we define the phonological information loss as

$$\mathcal{L}^{k}_{\text{phonology}} = |\mathcal{I}^{k\text{-Healthy}}_{\text{transm-posteriors}} - \mathcal{I}^{k\text{-pathology}}_{\text{transm-posteriors}} - \mathcal{L}^{k}_{\text{binary}}| \tag{8}$$

where $|.|$ stands for the absolute value. To obtain the phonological information loss $\mathcal{L}^{k}_{\text{phonology}}$, the difference in posterior information content is normalized by the binary difference obtained as

$$\mathcal{L}^{k}_{\text{binary}} = |\mathcal{I}^{k\text{-Healthy}}_{\text{transm-binary}} - \mathcal{I}^{k\text{-pathology}}_{\text{transm-binary}}| \tag{9}$$

If healthy and pathological speakers read different texts, this quantity is non-zero, so the effect of binary information difference between healthy and pathological speech is factored out in (8). If the phonetic transcriptions are the same, $\mathcal{L}^{k}_{\text{binary}} = 0$.

Applying a phoneme perception method operating on the principle of independent processing of compositional phonological classes provides a measure of the production influency that a listener may perceive. This idea is described in the following section IV-B.

B. Information Loss in Phoneme Perception

Initial works to understand human perception are conducted on recognition of syllables. A principle proposal of the studies pioneered by Harvey Fletcher is that humans appear to perform partial recognition of phonetic units in different frequency ranges independently [10], [11], [29], [30].

Recent studies by Nima Megarani and colleagues [9] suggest that phone attributes contain disjoint frequency components. The evidence is demonstrated as the weighted average spectro-temporal receptive fields (STRF) of the neural activities clustered on the phone attributes (cf. Fig. 2 of [9]). Hence, building

on Fletcher's proposal, we assume that phonetic perception relies on independent processing of multiple streams of phone attribute inference as depicted in Fig. 3.

Fig. 3. Phoneme decoding: Phoneme perception operates on the basis of merging evidences on phone attributes composition.

The goal is to assess speech production quality by drawing inference on the underlying phonemes using phonological posteriors. We define the l^{th} phonological composition for phoneme s_l as the set of K_l phonological classes, thus $g_{s_l} = \{q_1, \ldots, q_{K_l}\}$. The probability of erroneous phoneme perception is obtained as multiplication of the products of errors at individual phonological classes. Hence, the compositional probability of phoneme perception is expressed as

$$p(g_{s_l}, z_t) = 1 - (1 - p(q_1, z_t)) \ldots (1 - p(q_{K_l}, z_t)) \tag{10}$$

To obtain $p_{g_{s_l}}$, z_t is marginalized assuming a uniform probability for the available T_{s_l} frames aligned (using phonetic transcription) as the phoneme s_l via

$$p_{g_{s_l}} = \frac{1}{T_{s_l}} \sum_{t=1}^{T_{s_l}} p(s_l, z_t) \quad \forall l \in \{1, \ldots, L\} \tag{11}$$

That amounts to the phoneme uncertainty calculated as

$$\mathcal{H}^l_{\text{posteriors}} = -p_{g_{s_l}} \log_2 p_{g_{s_l}} \quad \forall l \in \{1, \ldots, L\} \tag{12}$$

The quantity $\mathcal{H}^l_{\text{posteriors}}$ determines the uncertainty pertained to perception of an individual phoneme by processing the inference of the phone attributes obtained in phonological posteriors.

If speech production is performed fluently, the phonological posteriors get close to their ideal binary values [31]. Due to a unique phonological composition defined for every phoneme, sharp posteriors lead to a minor uncertainty in phoneme perception. On the other hand, high uncertainty at the level of phone attributes (small posterior probabilities) leads to a great uncertainty in phoneme identification[1].

[1]We admit the fact that human perception may not operate at the level of phoneme classification or phonological composition. Hence, this approach is not a computerized version of quantifying human perception. Rather, it takes an initial step towards formalizing the speech perception as the process of decoding the speech code that corresponds to the phonological compositions in the present work. Nevertheless, alternative codes can be considered for information transmission analysis.

As a use case on pathological speech assessment, we are now able to quantify the level of information loss in phoneme perception based on the degradation in phonological posteriors. Therefore, we can find out which phonemes are affected the most due to impaired speech production.

To that end, we calculate $H^l_{\text{posteriors}}$ (12) and measure its distance to the information obtained from the binary (ideal) mapping. The phoneme information loss is thus defined as

$$\mathcal{L}^l_{\text{phoneme}} = |\mathcal{H}^l_{\text{posteriors}} - \mathcal{H}^l_{\text{binary}}| \tag{13}$$

where $\mathcal{H}^l_{\text{binary}} = H(s_l)$ as obtained from (2) using the phoneme probabilities estimated in Algorithm 1. The phonemes with larger distances from the binary canonical information are the ones whose perception are most distorted.

In the next section, we exploit the uncertainty pertained to phonemes for individual speakers. We propose a metric to assess production fluency with respect to an ideal speech production that can distinguish apart healthy and pathological speech.

C. Compositional Information Index

We assume that a speaker has produced T_{spk} speech frames, resulting in $\{p(g_{s_l}, z_t)\}_{t=1}^{T_{\text{spk}}}$ compositional phoneme probabilities corresponding to $\{g_{s_l}\}_{l=1}^{L}$ obtained from (10). Hence, the speaker-specific probabilities are estimated through marginalization over z_t assuming a uniform probability as

$$p^{\text{spk}}_{g_{s_l}} = \frac{1}{T_{\text{spk}}} \sum_{t=1}^{T_{\text{spk}}} p(g_{s_l}, z_t) \tag{14}$$

We define a compositional information (CI) index to assess perception of the production fluency expressed as

$$\text{CI} = \frac{\sum_{l=1}^{L} p^{\text{spk}}_{g_{s_l}} \log_2 p^{\text{spk}}_{g_{s_l}}}{\sum_{l=1}^{L} p^{\text{spk}}_{s_l} \log_2 p^{\text{spk}}_{s_l}} \tag{15}$$

where $p^{\text{spk}}_{s_l}$ indicates the phoneme probability for a speaker obtained by the frequency approach based on counting the phonemes in the speaker's phonetic transcription.

The CI index can be used to determine the degree of fluency in speech production exhibited in probabilistic phone attribute characterization with respect to the binary mapping in an ideal production (probabilities equal to 0 or 1). This score is expected to be small as the speech production is disrupted. We will see through the numerical evaluation in Section V that CI index enables separation of healthy and pathological speech with a large margin.

V. Numerical Evaluation

Numerical studies are conducted to demonstrate the potential of the proposed information theoretic framework to assess the quality of speech production based on the notion of information loss exploiting probabilistic characterization of the phone attributes.

A. Experimental Setup

1) Data: We use the WSJ database [32] to train the DNNs for phonological analysis. The training set was the 90% subset of the WSJ *si_tr_s_284* set, and the remaining 10% subset was used for cross-validation. The phoneme set comprises 40 phonemes (including "sil", representing silence) defined by the CMU pronunciation dictionary.

As evaluation data, we used the TORGO database of dysarthric speech that consists of recordings from speakers with either cerebral palsy or amyotrophic lateral sclerosis [33], along with Frenchay Dysarthria Assessment (FDA) [34] done by a speech-language pathologist. Original data include 3 female and 5 male pathological speakers, and 3 female and 4 male control (healthy) speakers. The recordings of dysarthric speech have been manually checked, and those with significant clipping waveform distortion have been removed from further analysis.

The Frenchay assessment includes 28 relevant perceptual dimensions of speech, namely related to the following dimensions:

- Laryngeal: noting weather the patient has clear phonation with the vocal folds, without huskiness.
- Tongue: noting accurate tongue movements (positions) with correct articulation.
- Palate: noting nasal resonance in spontaneous conversation, without hypernasality or nasal emission.
- Lips: observing the movements of lips in conversation, noting correct shape of lips.
- Respiration: noting running out of breath when speaking, and breathy voice.

2) Training: We use our open-source phonological vocoding platform [20] to perform phonological analysis. Briefly, the platform is based on cascaded speech analysis and synthesis that works internally with the phonological speech representation. In the phonological analysis part, phonological posteriors are extracted from the speech signal by DNNs. We used the binary classification of the eSPE set [19], and thus each DNN determines the probability of a particular phonological class.

To train the DNNs for phonological analysis, we first trained a phoneme-based automatic speech recognition system using Mel frequency cepstral coefficients (MFCC) as acoustic features. The three-state, cross-word triphone models were trained with the HMM-based speech synthesis system (HTS)

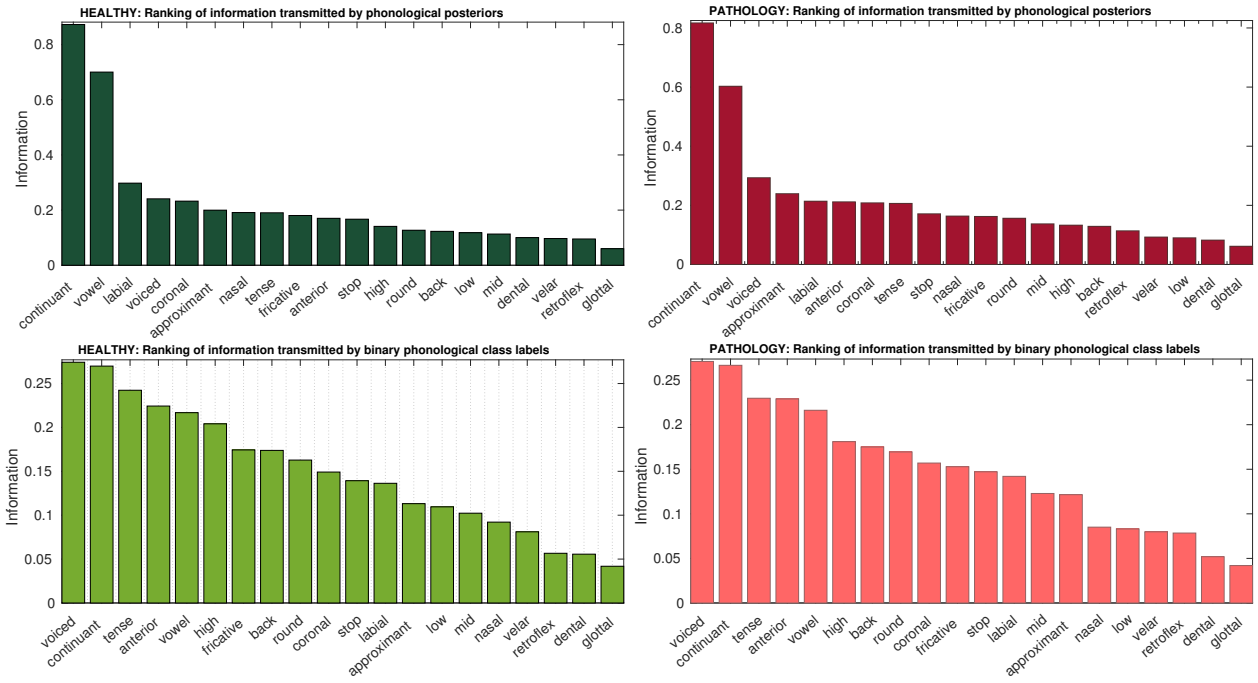

Fig. 4. Ranking of information content in phone attributes: The information quantities (bits) for probabilistic and binary phonetic-phonological associations are calculated from $\mathcal{I}^k_{\text{transm-posteriors}}$ (6) and $\mathcal{I}^k_{\text{transm-binary}}$ (5) respectively. We can see that the probabilistic estimation of information quantity shows a variance greater than the theoretical binary information. The top 4 most important phone attributes are identified as [continuant], [vowel], [labial], and [voiced].

variant [35] of the Hidden Markov Model Toolkit (HTK) on the WSJ training and cross-validation sets. The acoustic models were used to get boundaries of the phoneme labels, which were mapped to the eSPE phonological classes. In total, 21 DNNs were trained as phonological analyzers using the short segment (frame) alignment with two output labels indicating whether the phonological class exists for the aligned phoneme or not. In other words, the two DNN outputs correspond to the target class vs. the rest.

Each DNN was trained on the whole training set. The DNNs have an architecture of $351 \times 1024 \times 1024 \times 1024 \times 2$ neurons, determined empirically based on the authors' experience. The input vectors are 39 order MFCC features with the temporal context of 9 successive frames. The parameters were initialized using deep belief network pre-training following the single-step contrastive divergence (CD-1) procedure of [36]. The DNNs with the softmax output function were then trained using a mini-batch based stochastic gradient descent algorithm with the cross-entropy cost function of the KALDI toolkit [37]. The DNN outputs for individual phonological classes determine the phonological posterior probabilities. Detection accuracies of the eSPE phonological classes are very high (cf. Table III of [38]).

3) Phonetic Alignment: Evaluation data were aligned using the HTK tools, with the WSJ HMMs and the CMU dictionary [39]. Overall 6278 utterances were successfully processed, with 4374 recordings from the control speakers, and 1904 recordings from the speakers with dysarthria.

B. Ranking of Phonological Information

We calculate the information content of the phone attributes. The information can be quantified using the binary table of phonetic-phonological mapping as summarized in Algorithm 1. The binary maps used in this work are taken from Appendix A of [19]. The information content of an individual phone attribute corresponds to $\mathcal{I}^k_{\text{transm-binary}}$ (5). Alternatively, continuous phonological posteriors can be exploited to obtain $\mathcal{I}^k_{\text{transm-posteriors}}$ (6). The resulted mutual information is quantified for each frame. Considering a long duration of multiple frames, we compute an average of the mutual information.

The results are sorted and demonstrated in Fig. 4 for both healthy and pathological speech production. Comparing the binary and probabilistic information content indicates that the difference between highly informative attributes such as [continuant] or [vowel] and less informative attributes such as [glottal] and [dental] is far greater when their probabilities are inferred from the acoustic speech signal through phonological posteriors. This observation may indicate that some phone attributes make a higher impact on structure of the speech signal, and they bear more information in detection of the phoneme identities.

The ranking is different for impaired speech production implying that the information loss may not be equal for all phone attributes. In other words, speech production impairment may be more visible if a subset of phonological classes is selected [40].

C. Redundancy of Compositional Information

A phoneme identity is defined by composition of a few underlying phonological classes. We calculate the redundancy as the difference between constituting phonological information and the phoneme information as expressed in (7).

Fig. 5 shows the ranking of redundancy in production of phonemes. A difference is observed between two groups of the phonemes characterized by the vowels, and the stops and africates. The results imply that the latter consonantal group of the phonemes is less robust in the presence of distortion. Comparing the differences of the healthy controls and the speakers with dysarthria, the robustness to distortion is similar.

Although, the redundancy analysis suggests that a small subset of phonetic attributes may suffice to determine the phoneme categories, development of speech production involving redundancy [27], [28]

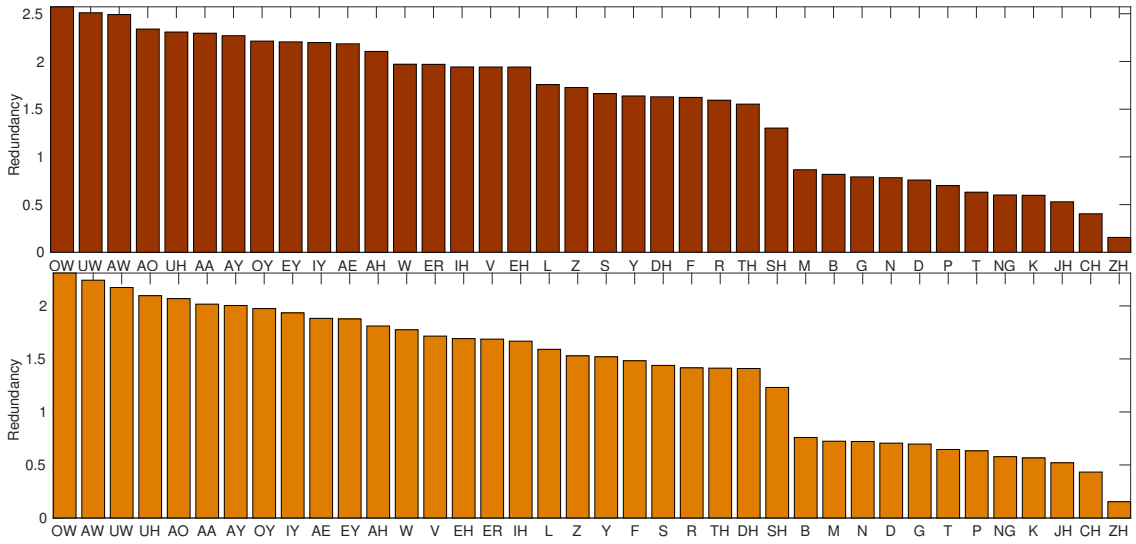

Fig. 5. Ranking of redundancy in production of phonemes as a combination of multiple phone attributes (7) for (top) healthy and (bottom) pathological speech production. Phonological posteriors are used to obtain the results illustrated above. Similar results are obtained for binary phonetic-phonological association. The phone description is according to [39].

TABLE I

CI index calculated from (15) is listed for each speaker. We can see that the healthy and pathological productions can be distinguished with a large margin.

Condition	**Healthy**						**Pathology**				
Speaker	FC01	FC02	FC03	MC01	MC02	MC03	F01-	F03-	M05-	M01-	M02-
CI (15)	2.91	2.47	2.10	8.06	3.89	2.58	0.25	0.21	0.51	0.18	0.20

may ensure robustness in adverse acoustic conditions.

D. Information Loss

Information loss is the difference between the information content of a phone attribute when it is obtained from healthy and pathological production. This quantity is calculated based on the expression in (8). The healthy and pathological speakers read different texts, thus the effect of different underlying phonetic transcriptions is normalized. Fig. 6 illustrates the information loss due to speech pathology.

The ranking of the phonological classes corresponds to the Frenchay assessment. The [continuant] and [vowel] classes correspond to the laryngeal dimension, where clear phonation is necessary to produce correct vowels, without any significant obstruction in the vocal tract. The [labial] class is associated with

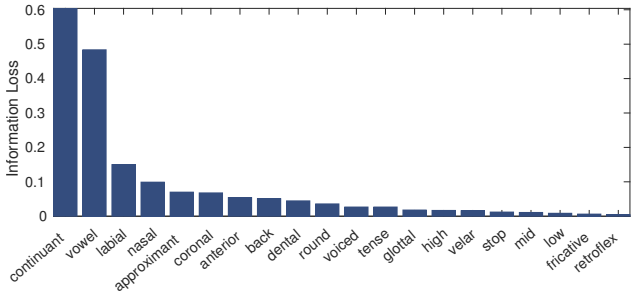

Fig. 6. Phone attribute information loss due to production impairment calculated based on (8).

the lips dimension and the [nasal] class with the palate dimension. The [coronal] and [anterior] classes are related to the tongue dimension, where the former is related to the tongue-tip, and the latter to the tongue-blade.

To quantify the effect of phone attribute information loss on phoneme perception, we apply the method explained in Section IV-C. The quantity of information loss in phoneme perception is calculated from (13). The results are illustrated in Fig. 7 for the top 20 phoneme categories affected by speech pathology.

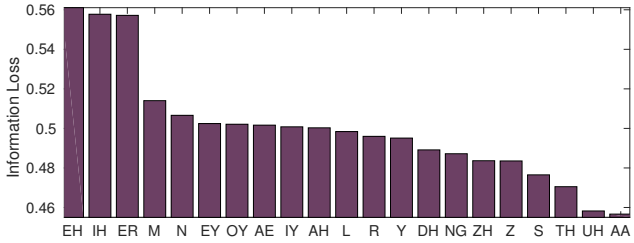

Fig. 7. Phoneme perception information loss due to production impairment calculated based on (13) and demonstrated for the top 20 most affected phonemes. The phonemes are described in [39].

This observation suggests that the effect of impaired production in perception of phonemes is not equally distributed and investigations on a selected category of phonemes may bring practical benefits in assessment of pathological speech production.

The group of the first three top most affected vowels refers to the high-front and the mid-central phonemes that might be associated with the Tongue dimension of the Frenchay assessment.

E. Detection of Pathological Speech

Finally, we evaluate the proposed CI index for both cases of healthy and pathological speech. Building on our observation on ranking of the influence of speech production disorder in phoneme categories, CI

is calculated for phoneme /EH/ which shows the greatest effect. The results are listed in Table I. We can see that the scores of pathological speech are small (as expected) and they are distinguished from the healthy CI by a large margin. In this analysis, we provide a single measure for the whole speech data of each speaker, and the minimum length of the data sufficient for detection remains to be studied in our future work.

VI. Concluding Remarks

An information theoretic analysis of speech production and perception is proposed exploiting probabilistic characterization of the phone attributes using DNNs. The resulted framework quantifies the quality of speech production and measures the amount of information loss due to production inaccuracy. The information loss in phone attributes enables us to quantify the measure of information loss in perception of phonemes defined as a composition of phone attributes. In this context, variations is speech production can be compared and contrasted.

As a case study, we evaluate the proposed method for assessment of information loss due to production impairment in speech pathology. A novel compositional information (CI) index is defined as the ratio of speaker's production information and its information in ideal production. The CI scores low for pathological production and enables us to distinguish the cases of speech pathology in the TORGO database from the control healthy speakers.

The applications of this analysis approach may be far beyond in other domains relying on DNN posterior probabilities such as speech recognition, speech coding, spoken query detection, as well as language (pronunciation) learning. This framework makes it possible to find the elements of information loss and degradation through transmission analysis of application-specific channels. It also paves the way for quantitative and computerized evaluation of neuro-linguistics and phsyco-linguistics experiments.

Acknowledgment

Afsaneh Asaei has been supported by SNSF project on "Parsimonious Hierarchical Automatic Speech Recognition (PHASER)" and PHASER-QUAD grant agreement numbers 200021-153507, 200020-169398. We acknowledge Prof. Marina Laganaro from university of Geneva for her valuable comments on this paper.

References

[1] N. Dronkers and J. Ogar, "Brain areas involved in speech production," *Brain*, vol. 127, no. 7, pp. 1461–1462, Jul. 2004.

[2] W. J. Levelt, *Speaking: From Intention to Articulation*. ACL-MIT Press Series in Natural-Language Processing, 1989.

[3] W. J. Levelt, A. Roelofs, and A. S. Meyer, "A theory of lexical access in speech production." *The Behavioral and brain sciences*, vol. 22, no. 1, Feb. 1999.

[4] C. P. Browman and L. M. Goldstein, "Towards an articulatory phonology," *Phonology*, vol. 3, pp. 219–252, 1986.

[5] ——, "Articulatory gestures as phonological units," *Phonology*, vol. 6, pp. 201–251, 1989.

[6] ——, "Articulatory phonology: An overview," *Phonetica*, vol. 49, pp. 155–180, 1992.

[7] K. E. Bouchard, N. Mesgarani, K. Johnson, and E. F. Chang, "Functional organization of human sensorimotor cortex for speech articulation." *Nature*, vol. 495, no. 7441, pp. 327–332, 2013.

[8] F. H. Guenther and G. Hickok, *Role of the auditory system in speech production*. Elsevier, 2015, vol. 129, pp. 161–175.

[9] N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang, "Phonetic feature encoding in human superior temporal gyrus," *Science*, vol. 343, no. 6174, pp. 1006–1010, 2014.

[10] B. Gold, N. Morgan, and D. Ellis, *Speech and audio signal processing: processing and perception of speech and music*. John Wiley & Sons, 2011.

[11] H. Bourlard and S. Dupont, "A new ASR approach based on independent processing and recombination of partial frequency bands," in *International Conference on Spoken Language (ICSLP)*, 1996, pp. 426–429.

[12] R. L. Diehl, A. J. Lotto, and L. L. Holt, "Speech perception," *Annu. Rev. Psychol.*, vol. 55, pp. 149–179, 2004.

[13] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, vol. 21, no. 1, pp. 1–36, 1985.

[14] C. A. Fowler, D. Shankweiler, and M. Studdert-Kennedy, "Perception of the speech code revisited: Speech is alphabetic after all," *Psychological Review*, vol. 123(2), pp. 125–150, 2015.

[15] A. M. Liberman and I. G. Mattingly, "The motor theory of speech perception revised," *Cognition*, pp. 1–36, 1985.

[16] C. P. Brownian and L. Goldsteint, "Gestural specification using dynamically-defined articulatory structures," *Journal of Phonetics*, vol. 18, pp. 299–320, 1990.

[17] N. Chomsky and M. Halle, *The Sound Pattern of English*. Harper & Row, 1968.

[18] D. Yu, S. M. Siniscalchi, L. Deng, and C.-H. Lee, "Boosting attribute and phone estimation accuracies with deep neural networks for detection-based speech recognition," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2012, pp. 4169–4172.

[19] M. Cernak, S. Benus, and A. Lazaridis, "Speech vocoding for laboratory phonology," *Computer, Speech and Language*, vol. 42, pp. 100–121, 2017.

[20] M. Cernak and P. N. Garner, "PhonVoc: A Phonetic and Phonological Vocoding Toolkit," in *Proc. of Interspeech*, 2016.

[21] G. A. Miller and P. E. Nicely, "An analysis of perceptual confusions among some english consonants," *The Journal of the Acoustical Society of America*, vol. 27, no. 2, pp. 338–352, 1955.

[22] D. J. Oosthuizen and J. J. Hanekom, "Information transmission analysis for continuous speech features," *Speech Communication*, vol. 82, pp. 53–66, 2016.

[23] M. Cernak, A. Asaei, and H. Bourlard, "On structured sparsity of phonological posteriors for linguistic parsing," *Speech Communication*, vol. 84, pp. 36–45, 2016.

[24] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[25] R. Swanepoel, D. J. Oosthuizen, and J. J. Hanekom, "The relative importance of spectral cues for vowel recognition in severe noise," *The Journal of the Acoustical Society of America*, vol. 132, no. 4, pp. 2652–2662, 2012.

[26] N. Timme, W. Alford, B. Flecker, and J. M. Beggs, "Synergy, redundancy, and multivariate information measures: an experimentalists perspective," *Journal of computational neuroscience*, vol. 36, no. 2, pp. 119–140, 2014.

[27] G. Hickok and D. Poeppel, "The cortical organization of speech processing," *Nature Reviews Neuroscience*, vol. 8, no. 5, pp. 393–402, 2007.

[28] G. Hickok, "The cortical organization of speech processing: Feedback control and predictive coding the context of a dual-stream model," *Journal of communication disorders*, vol. 45, no. 6, pp. 393–402, 2012.

[29] J. B. Allen, "How do humans process and recognize speech?" *IEEE Transactions on speech and audio processing*, vol. 2, no. 4, pp. 567–577, 1994.

[30] R. P. Lippmann, "Speech recognition by machines and humans," *Speech communication*, vol. 22, no. 1, pp. 1–15, 1997.

[31] A. Asaei, M. Cernak, and H. Bourlard, "On Compressibility of Neural Network Phonological Features for Low Bit Rate Speech Coding," in *Proc. of Interspeech*, 2015, pp. 418–422.

[32] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, ser. HLT '91. Association for Computational Linguistics, 1992, pp. 357–362.

[33] F. Rudzicz, A. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[34] P. M. Enderby, *Frenchay dysarthria assessment*. College Hill Press, 1983.

[35] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based Speech Synthesis System Version 2.0," in *Proc. of ISCA SSW6*, 2007, pp. 131–136.

[36] G. E. Hinton, S. Osindero, and Y. W. Teh, "A Fast Learning Algorithm for Deep Belief Nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[37] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, "The kaldi speech recognition toolkit," in *Proc. of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2011.

[38] M. Cernak, A. Lazaridis, A. Asaei, and P. N. Garner, "Composition of Deep and Spiking Neural Networks for Very Low Bit Rate Speech Coding," *IEEE/ACM Trans. on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2301–2312, 2016.

[39] R. L. Weide, "The CMU pronouncing dictionary," 1998. [Online]. Available: http://www.speech.cs.cmu.edu/cgi-bin/cmudict

[40] A. Asaei, M. Cernak, and M. Laganaro, "PAoS markers: Trajectory analysis of selective phonological posteriors for assessment of progressive apraxia of speech," in *Proceeding on the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016.